International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

A Reliable Method for Accuracy Constrained Privacy Preservation for Relational Data

Waghmare Ashwini R[#1], V. M. Jarali[*2]

MBES's College of Engineering, Ambajogai, Dr. B.A.M.U

Abstract: ***Today Information is most likely the vital and demanded resource. Distribution, transfer, mining and publishing data are primary operations in everyday life. Preserving the privacy of persons is indispensable one. Sensitive private information must be protected when data is published. The data should be uncovered through proper access control mechanism.Unauthorized data observation results in the disclosure of information to users not authorized to gain access to such information.There are two kinds of risks namely identity disclosure and attribute disclosure that affects privacy of persons whose data are published. Premature Researchers have contributed new methods specifically k-anonymity, l-diversity to protect privacy. K-anonymity method preserves privacy of persons against identity disclosure attack only. But Attribute disclosure attack makes negotiation this method. Limitation of k-anonymity is satisfied through l-diversity method. But it does not satisfy the privacy against the two attacks those are identity disclosure attack and attribute disclosure attack in several scenarios. To preserve the privacy of individuals' sensitive information from attribute and identity disclosure attacks a new method is proposed. This method minimization information loss and gains the privacy by using generalization algorithm which is proposed in this method and is described in this paper.***

Keywords: Privacy Preservation, Data mining, Anonymization, k-anonymity, access control

1. Introduction

Today Information is most likely the vital and demanded resource.The internetworked society that relies on the distribution and sharing of information in the private as well as in the public and governmental sectors. Governmental, public, and private institutions are gradually more required to make their data electronically available. If in the past this distribution and sharing of information was mostly in statistical and tabular form, many situations require today that the specific stored data themselves, called microdata, be released. The advantage of releasing microdata instead of specific pre-computed statistics is an improved flexibility and ease of use of information for the users [2]. Organizations collect and analyse customer data to get better their services. Access Control Mechanisms are used to make sure that only authorized information is available to users. However, sensitive information can still be misused by authorized users to compromise the privacy of consumers. Unauthorized data observation results in the disclosure of information to users not entitled to gain access to such information [6]. A Privacy Protection Mechanism (PPM) is not in place and we are sharing sensitive information then an authorized user can still compromise the privacy of a person which may lead to identity disclosure [1].

The anonymity techniques can be used with an access control mechanism to make sure both security and privacy of the sensitive information. Anonymization algorithms use suppression and generalization of records to satisfy privacy requirements with minimal distortion of resealed data or micro data. The anonymity techniques can be used with an access control mechanism to be sure about both security and privacy of the sensitive information of the data. In this system the focus is only on a static relational table which is anonymized only once. However, the concept of accuracy constraints for permissions can be useful to any privacy-preserving security policy, e.g., discretionary access control.

2. Related Work

Data anonymization is nothing but preventing leakage of a data to others. If data is released then there is high possibility of misuse of data. So that the privacy of individuals can get compromised. However, in the concept of anonymity, a rule applying to a small unit of individuals records gives a more serious threat because record owners from a small group are more identifiable [9]. There are many existing techniques for data anonymization such as *k*-anonymity, *l*-diversity. But k-anonymity[4][7][8] is prone to the homogeneity attack and the background knowledge attack, but it does not attain the privacy preservation against attribute disclosure attack but can attain for identity disclosure attack. *l*-diversity [4] overcomes the weakness of k-anonymity and satisfy the privacy preservation against attribute disclosure attack. But, its efficiency is not superior in case of identity disclosure attack. *l*-diversity method overcomes the drawbacks of k-anonymity. In this method, an equivalence class is said to have *l*-diversity if there are at least *l* "well-represented" values for the sensitive attribute. A table is said to enclose ℓ-diversity if every equivalence class of the table has ℓ-diversity. For sensitive numeric attributes, an *l*-diverse equivalence class can still disclose information if the numeric values are close to each other.

In t-closeness method [13], an equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to satisfy t-closeness if all equivalence classes have t-closeness. P sensitive k-anonymity model, the micro data table T* satisfies (p+, α)-sensitive k-anonymity property if it satisfies k-anonymity[10]. Tamir Tassa [11] proposed an alternative model of k-type anonymity. It reduces the information loss than k-anonymity and obtain anonymized table by less generalization. It preserves the privacy against identity disclosure alone. Qian Wang[12]has proposed a model for kanonymity in security of attribute disclosure. It

can prevent attribute disclosure by controlling average leakage probability and probability difference of sensitive attribute value.

3. Proposed System

A new anonymization method is proposed here which helps to improve the security and data quality. It works against attribute disclosure and identity disclosure and provides security to sensitive data and also preserves privacy of sensitive attributes. The technique cannot guarantee privacy which includes some operations like removing the unique identifiers such as Name or Id from the table. Sometimes these approaches can also leak sensitive information about an individual. Other Quasi attributes like Date of Birth, Sex, PIN Code when combined together from other published data can also reveal the identity of an individual.

If two different data sets are published and one containing some attribute values which are also present in other one then Re-identification is possible by combining those attribute values. The existing anonymity techniques have some drawbacks which are related to information disclosure such as attribute disclosure and identity disclosure and also there is some lack of privacy for individual attributes in a Database. The proposed method overcomes these problems and provides privacy to individuals and reduces the limitations of existing anonymity techniques with privacy in new mode.

Proposed method includes two steps. First of all it performs generalization operation with selected quasi identifiers QI and then generate anonymization table T. This technique described in detail as follows. The records in sensitive table T are placed and in units U_1,U_2,U_3, …..U_n by applying generalization. Proposed algorithm is applied on each unit.Consider a_j is the lower limit of the generalization range and b_j is the upper limit of generalization range. The value of a_j is set as the current value of numeral quasi identifier and b_j is the next maximum value of a_j. If b_j is not found in any group then $\leq a_j$ this condition is set. The working of algorithm is given below:

Algorithm: Anonymization

Input: T
Output: T*
Step1 : Start

Step2 : Make the units $U_1,U_2,U_3….U_{n-1},U_n$ of the records in T according to Quasi Identifier QI_i
Where, i= 1,2,3,……,k.
Step3 : **for**($U_k \in T$) **do**
Step4 : Sort each record in ascending order in order to QI
Step5 : let $a_i = val[QI]_i$
Step6 : **for**($a_i \in U_k$) **do**
Step7 : Find next maximum value b_i of a_i
Step8 : **If** (b_i is found) **then**
Step9 : Generalization condition = $[a_i-b_i]$
Step10: **else**
Step11: Generalization condition = $\leq a_i$
Step12: End

Above proposed algorithm is very efficient in case of attribute disclosure and identity disclosure because it sets the minimal generalization range so than less information loss occurs. It also gives more privacy gain compared with existing techniques like *k*- anonymity and *l*-diversity.

The calculation of information loss and privacy gain is given in section V.

4. Implementation

The implementation contains the modules which are given in fig.1. Working of each module is explained below:

- *Admin:* In this module, the Admin has to login by using applicable user name and password. After successfully login, he can do some operations such as search users, query cut, median cut, view list users, view attackers, data recovery and then logout.
- *Search Users:* In this module, the admin can see two types of data table: 1) sensitive data and 2)anonymous data. In sensitive data, he can see the particular disease, pin code, age and Id of all registered users. In anonymous data, he can see the diseases between ages(e.g.: 0-10) and pin codes (e.g.: 40-60). In this system we are hiding the information about patient details and showing the anonymous records of patients.

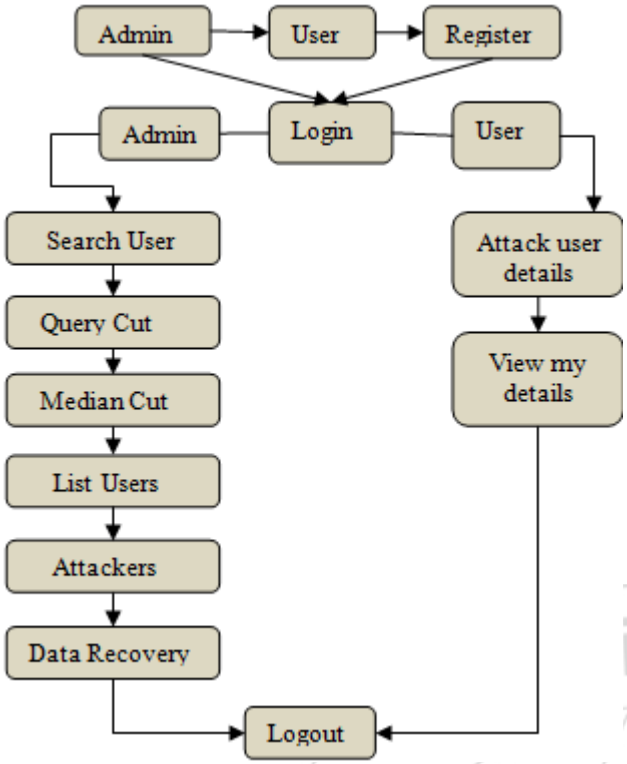

Figure 1: Abstract Diagram

- *Query Cut:* In this module, the admin can look for the diseases details based on the key words such as age and disease, then the server will search the record details connected to key words, then response will send to particular user.
- *Median cut:* In this module, the admin can look for the diseases based on the age and blood group, then server will mine the all data and send the linked data to particular user.
- *List of users:* In this module, the Admin can see list of all registered users. If the admin clicks on users button, then it will show all registered users with their tags such as user ID, user name, blood group, diseases, E mail ID, mobile no, Location, date of birth, address and pin code.
- *Data Recovery:* In this module, the admin can recover the modified data of attacked data. When any authorized user attack a data the admin will recover the attacked data and again upload to the database. So the accuracy of original data gets maintained.
- *User:* In this module, there are N numbers of users are present. Users have to register first. After registering successfully he has to login by using authorized user name and password. If he Login successfully, he can do some operations like attack user details, view my details and logout. If user want to see his own details, he has to clicks on my details button, then the server will give response to the user with their tags such as user ID, name, mobile no, address, pin code and email ID. If that authorised user wants to attack the particular user information and modify data, then he will click on attack user details button, then enter user name to attack particular user and submit. The server will display the user details of that particular user, and then he can edit the user information then click submit and server will give response to user. After modifying that data, then user will be considered as an attacker. And his name goes to the attacker details which will be stored in an attacker module.

5. Results and Discussion

The performance of proposed anonymization technique calculated in terms of two data metrics namely information loss and privacy gain [5].

The following formula calculates the information loss for generalizing v to v*:

$$Information\ Loss\ of\ (v^*) = \frac{(The\ number\ of\ values\ in\ v\ ^*)}{(The\ no\ of\ values\ in\ domain\ A)}$$

Where,
V is the value of domain of attribute A.
If Information Loss of (v*) = 0, then can be said that if v is an original data value in the table.

The loss of a generalized record r is given by

$$Information\ Loss\ (T^*) = \sum_{\forall t^* \epsilon T^*} Information\ Loss\ of\ (T^*)$$

$$Privacy\ Gain = A(QId_j) - A_g(QId_j)$$

where ,$A(QId_j)$ denote the anonymity before applying generalization and $A_g(QId_j)$ indicate the anonymity after the performing generalization g.

$$ILPG(gen) = \frac{Information\ Loss\ of\ (T^*)}{Privacy\ Gain\ of\ (T^*)}$$

After evaluating the results we get the following table:

Table 1: Results

Methods	Information Loss	Privacy Gain	ILPG
k anonymity	5.496	8	0.687
l-diversity	8.50	8	1.0625
Proposed method	2.711	11	0.251

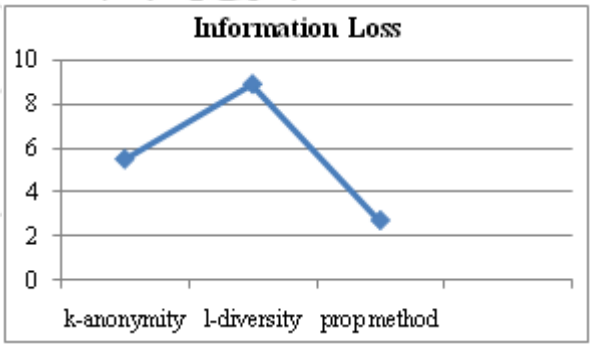

Figure 2: Measure of Information loss

When we calculate the information loss of k anonymity, l-diversity and proposed method we can observe that the proposed method have less information loss as compared to the other two methods as shown in fig. 2.

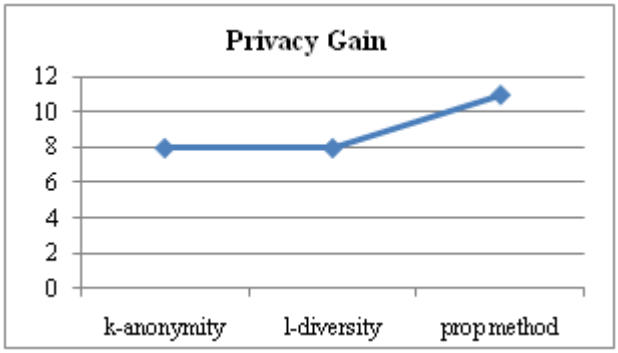

Figure 3: Measure of Privacy gain

The above graph in fig.3 is plotted for Privacy gain results of k-anonymity, l diversity and proposed method. As we can observe that proposed method provides more privacy to data as compared to the other two methods.

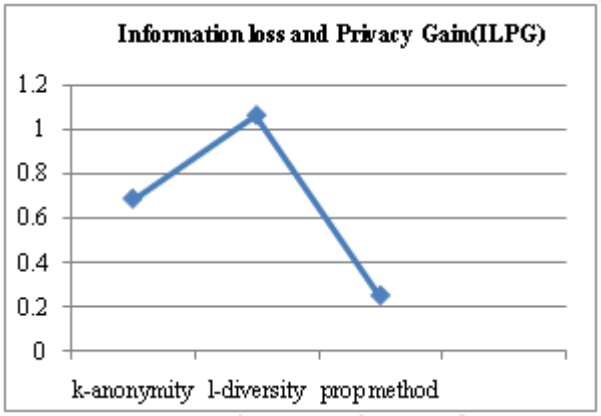

Figure 4: Measure of ILPG

It is also observed that the proposed method performs well in terms of privacy gain and ratio of information loss to privacy gain (ILPG).

6. Conclusion

Data transferring, sharing and publishing are growing in every day across the world. The usage of internet has increased and huge amount of data is generated and managed in every second, half of total volume of data contains sensitive information which requires to be managed with data security mechanism. In this situation, to pass proper authentication user needs to share different identity disclosure parameters which also are maintained as generalized data. In this paper a method is proposed which applied on static data. And the algorithm is applied on numeric quasi-identifiers. It provides the privacy and reduces the information loss. The proposed method achieves the preservation of individuals' sensitive information in anonymized Database. Performance of the proposed method is compared with existing methods in terms of privacy gain and information loss metrics. Results are tabulated and plotted. The results show improvement in privacy gain and reduction in information loss compared to existing methods. The proposed method is achieved the preservation of individuals' sensitive information which is represent as numeric values.

References

[1] Zahid Pervaiz, Walid G. Aref and Nagabhushana Prabhu "Accuracy-Constrained Privacy-Preserving Access Control Mechanism for Relational Data," IEEE transaction on Knowledge and data engg., vol.26, no. 4, april 2014
[2] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6,pp. 1010-1027, Nov. 2001.
[3] B. Fung, K. Wang, R. Chen, and P. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Surveys, vol. 42, no. 4, article 14, 2010.
[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-Diversity: Privacy Beyond k-anonymity," ACM Trans. Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.
[5] G. Ghinita, P. Karras, P. Kalnis, and N. Mamoulis, "Fast Data Anonymization with Low Information Loss," Proc. 33rd Int'l Conf. Very Large Data Bases, pp. 758-769, 2007.
[6] E. Bertino and R. Sandhu, "Database Security-Concepts, Approaches, and Challenges," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, Jan.-Mar. 2005.
[7] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional K-Anonymity," Proc. 22nd Int'l Conf. Data Eng., pp. 25-25, 2006.
[8] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A. Zhu, "Approximation Algorithms for k-Anonymity," J. Privacy Technology, vol. 2005112001, pp. 1-18, 2005.
[9] VERYKIOS, V. S.,ELMAGARMID, A. K.,BERTINO, E., SAYGIN, Y., AND DASSENI, E. 2004. Association rule hiding. *IEEE Trans. Knowl. Data Engin. 16*, 4, 434–447.
[10] Xiaoxun Sun, Hua Wang, Jiuyong Li and Traian Marius Truta, "Enhanced P-Sensitive K-Anonymity Models for privacy Preserving Data Publishing", Transactions On Data Privacy, 2008,pp53-66
[11] Tamir Tassa,Arnon Mazza and Aristides Gionis,"k-Concealment: An Alternative Model of k-Type Anonymity", TRANSACTIONS ON DATA PRIVACY 5,2012, pp189–222
[12] Qiang Wang,Zhiwei Xu and Shengzhi Qu,"An Enhanced KAnonymity Model against Homogeneity Attack", Journal of software,2011, Vol. 6, No.10, October 2011;1945-1952
[13] "t-Closeness: Privacy beyond k-Anonymity and l-Diversity" paper by Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian presentation by Caitlin Lustig for CS 295D